DOCUMENT RESUME

ED 037 713 AL 002 338

AUTHOR Uskup, Frances Land
TITLE A Method for Automating Dialect Analysis.
PUB DATE [70]
NOTE 13p.

EDRS PRICE EDRS Price MF-$0.25 HC-$0.75
DESCRIPTORS Codification, *Computational Linguistics, Computer Programs, *Data Processing, *Dialect Studies, Phonetic Transcription, Phonology

ABSTRACT

This paper proposes a method of handling limited problems in dialect research. In approaching the problem, it was necessary to devise a system for coding phonetic transcription which would take into account the variance in the diacritics of different field workers so that none of the material would be lost while permitting computer analysis. The design of the program also allows the researcher to isolate the significant variables found in the dialects examined. The author presents the coding system, the program organization and deck assembly instructions, a listing of the program and all the subroutines, and the informant coding. An accompanying computer print-out is available for inspection at the ERIC Clearinghouse for Linguistics, 1717 Massachusetts Avenue, N.W., Washington, D.C. 20036. Copies of the print-out are also available from the author at the Illinois Institute of Technology, Chicago, Illinois 60616. (Author/DO)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED037713

A METHOD FOR AUTOMATING DIALECT ANALYSIS

Frances Land Uskup

Dialect study has long been plagued with cumbersome analytical techniques. Over the past years, Linguistic Atlas data has slowly accumulated in researchers' file cabinets while attempts to analyze the great bulk of the material have languished for lack of appropriate techniques. Automation of data processing in dialect research, as in other fields, would seem to be the obvious solution, but there has been a general reluctance to implement this type of innovation except in limited cases.[1]

The main reason being the fear often expressed, that much of the fine phonetic detail would be lost in coding phonetic transcription. Another problem is that the machine print-out of data is a mixture of alfa-numeric characters when a coding system is used, which renders it useless for publication. It is not possible with the equipment available to most researchers to produce a list manuscript of items, except in a coded form. There is no company which offers a type-ribbon

---

[1] Shuy, R., Wolfram, W. A., and Riley, W. K., Field Techniques in an Urban Language Study. Center for Applied Linguistics: Washington, D.C. (1968).

AL 002 338

which contains IPA or Linguistic Atlas symbol notation, or for that matter, any combination of symbols which could be modified for such use. This "hardware problem" could be solved by a sizable research grant which would permit such equipment to be produced and installed.

This paper proposes a method of handling limited problems in dialect research. The method used will be presented as follows:

(1) Coding system

(2) Program organization and deck assembly instructions

(3) Listing of the program and all the subroutines

(4) Informant coding

In approaching the problem, it has been necessary to devise a system for coding phonetic transcription which would take into account the variance in the diacritics of different field workers so that none of the material would be lost while permitting computer analysis. Also, the design of a general sort program which would allow the researcher to isolate the significant variables found in the dialects examined has been accounted for.

The program is extremely versatile since it allows for extensive manipulation and sorting of the data. Any corpus which can be coded by the outlined system can be analyzed by this program. It would permit any researcher doing

dialect studies to analyze cheaply and quickly, large quanities of phonetic data. The coding system was designed for coding phonetic English transcription, but can be easily modified for phonemic analysis or for another language.

MACHINE CODE

Vowels	i	U01	ɘ	U12	ɣ	U23
	ɪ	U02	ɵ	U13	θ	U24
	e	U03	ɤ	U14	ʌ	U25
	ɛ	U04	ɜ	U15	ɑ	U26
	æ	U05	ɞ	U16	ɒ	U27
	a	U06	ɑ	U17	ɔ	U28
	y	U07	ɐ	U18	u	U29
	ʏ	U08	ɷ	U19	ʊ	U30
	ɸ	U09	ɤ	U20	o	U31
	ɨ	U10	ɞ	U21		
	ɨ	U11	ɯ	U22		

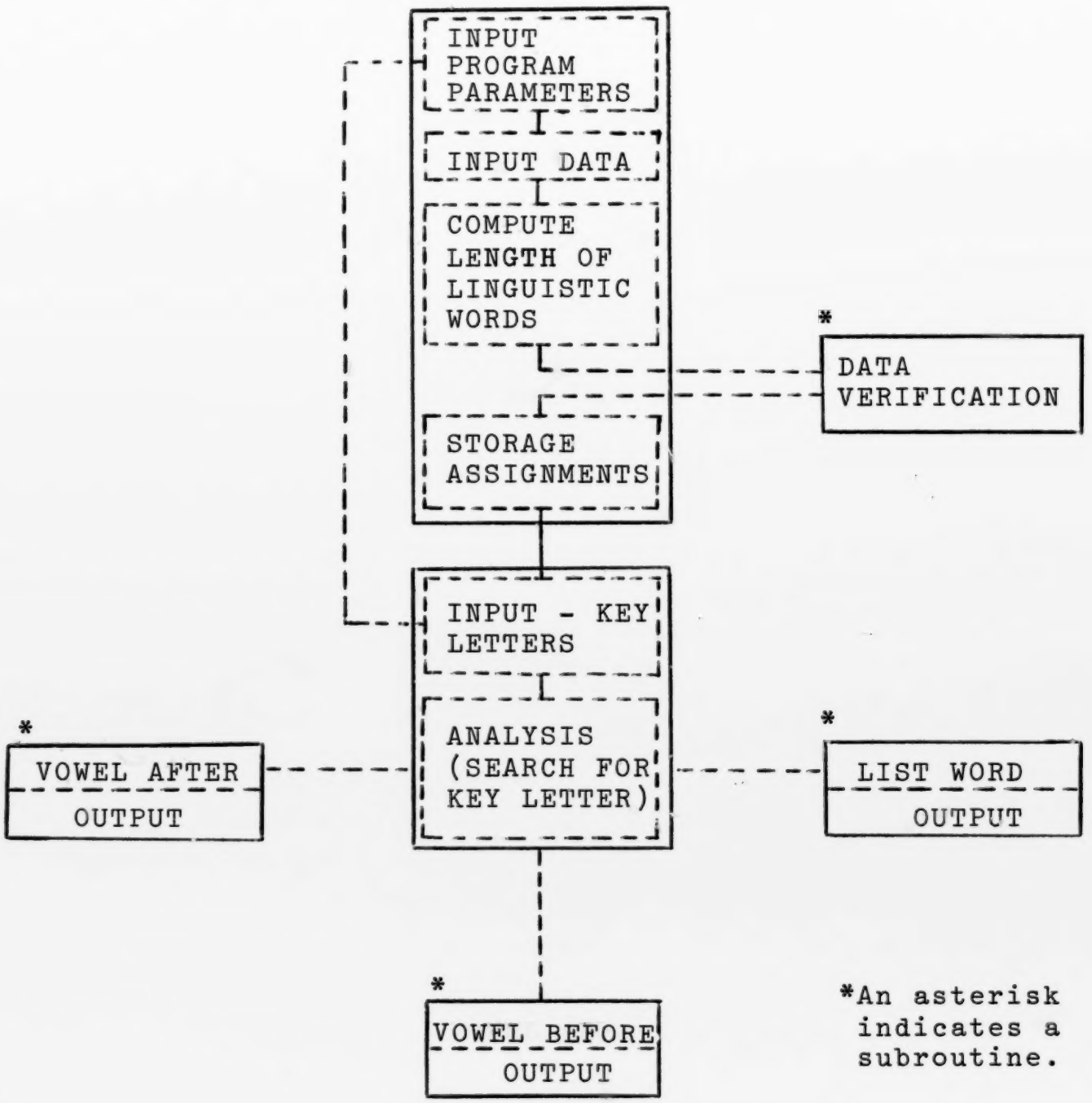
Consonants	b	B		l,	LX	s̆	SX
	B	BX		ɬ	LL	t	T
	β	BB		m	M	θ	TT
	č	C		n	N	v	V
	d	D		y	NX	z	Z
	ð	DD		p	P	ž	ZX
	f	F		ɸ	PP	w	W
	g	G		ʔ	Q	y	Y
	h	H		r	R	x	XX
	ǰ	J		ɾ	R$	j	JJ
	k	K		R	RX		
	l	L		s	S		

NOTE: Additional symbols can be added at any time to the system by combining any of the above symbols, or in the case of vowels, by adding additional numbers.

Vowel Modification	$V^{\wedge}$	raised	I01
	V^{v}	lowered	I02
	V<	fronted	I03
	V>	backed	I04
	$V_{\tilde{u}}$	nasalized	I05
	$V_{\tilde{l}}$	weakly nasalized	I06
	V	rounded unrounded	I07
	V´	slightly rounded	I08
	V.	length	I09
	V:	extra length	I10
	$V_{\text{ʕ}}$	laryngealized	I11
	$V_{\text{ʢ}}$	pharyngealized	I12
	$V_{\text{ɤ}}$	breathy	I13
	V^{v}	glide	I14
		example: a^{u} ----	U17I14U29
Consonant Modification	$C_{\text{ʇ}}$	coarticulation	?+
	$C_{\text{ʂ}}$	retroflex	$
	$C_{\text{ɔ}}$	syllabic	A
	C'	unreleased	+
	C_{x}	devoiced	,
	C<	fronted	*
	C	voiced devoiced	=
	C	voiceless voiced	'
	$C_{\text{͡}}$	lenis	8/5
	C>	backed	%
	C_{w}	rounded	#
	C'	aspiration	@
	Ċ	dentalized	1/4

PROGRAM ORGANIZATION

Shown below is a simplified view of the overall program structure.



The variables are referred to as "Key Letter(s)" or the "Sorting Base". When a valid Key Letter has been found in the linguistic word, either the entire word is listed, or the vowel (+ modification) that is present, before or after the key letter, is listed.

SUBROUTINES

The program has four subroutines which can be used in various ways, separately or in combination, depending on the needs of the researcher. The data can simply be listed for storage (2) or checked for errors in key-punching (1). Subroutines (3) and (4) make it possible to search the data and have the environment surrounding certain features one wishes to isolate listed along with item number and informant number. For example, one can list the vowels (± modifications) occurring before or after the variable to be analyzed. The subroutines will be listed below with a brief explanation of their function:

(1) Data Verification This task is performed in the subroutine VRFY. This subroutine checks a data card for the language code and informant number (for a one-informant group) to verify that it belongs with the language/dialect group of the informant set under analysis. Any erroneous data will be noted as such and listed, but it will not be processed in the analysis section of the program.

(2) List Word The subroutine LSTWD lists the pertinent information on the linguistic word that contains the key letter. The output can be from the printer or card punch, or both. The printer will list the page, the item number and the full linguistic word. The card punch duplicates the input card (language/dialect code, informant number, page number, item number and linguistic word).

(3) Vowel Before The subroutine VØWB, searches for a vowel or a vowel plus modification before the key letter(s) based upon an initial parameter selection. If a vowel is present, the output will list the page, item number and vowel (plus modification). The absence of vowel in that position as well as lack of modification will be noted on the output.

(4) Vowel After The subroutine, VØWA, searches for a vowel or a vowel plus modification after the key letter(s) based on the initial parameters selected. If a vowel occurs, the output will list the page number, item number and the vowel plus modification. The absence of vowel plus modification will be listed on the output.

PROGRAM INPUT

The program input consists of the source program and the following:

Control Card
Parameter Card
Data Identification Card
Linguistic Word Data Cards
Key Letter Data Cards

(1) Control Card Only one card is required per computer run.

Col. 1-2 Scratch file unit number, IUNIT. This variable designates which disk or tape unit is to be used for temporary storage.

NOTE: This number is assigned by the computer center.

Col. 3-4 An arbitrary constant, CDAT. This is used to indicate that another data set is to be processed. The value of CDAT can be any two digit number from 30 to 50.

An input format of 2I2 is used for the Control Card.
Sample:

(2) Parameter Card At least one card is required per computer run.

Col. 1 Parameter Continuation Variable, DFPAR. The value of this variable dictates whether or not the input section of the main program will be used.

A blank or zero - input new linguistic word data.
A 1 thru 9 - bypass input section and go directly to analysis section; use previous data.

Col. 2 — Designation of the type of data to be processed, NAL.

A 1 designates a single informant.
A 2 designates a group analysis.

Col. 3 — NEXT designates which subroutine is to be called after the Key Letter has been found in the linguistic word.

A 1 calls subroutine LSTWD.
A 2 calls subroutine VOWB.
A 3 calls subroutine VOWA.

Col. 4 — IVB designates whether the vowel or the vowel plus modification will be searched for in the subroutines VOWB and VOWA.

A 1 designates vowel + modification.
A 2 designates vowel only.

Col. 5 — NVR designates whether or not the data verification subroutine is called.

A 1 calls subroutine VRFY.
A 2 bypasses subroutine VRFY.

Col. 6 — NPNCH designates the type of output for subroutines LSTWD.

A 1 designates printer only.
A 2 designates card punch only.
A 3 designates printer and card punch.

Col. 7-12 — Must be blanks for proper execution of the program.

An input formant of 6I1, A4, I2 is used for the Parameter Card.
Sample:

(3) Data Identification Card One data identification card must precede the Linguistic Word Data Cards.

For Group Data:

Col. 1-2 Language code.

Col. 3-6 Group number.

For Single Informant Data:

Col. 1-2 Language code.

Col. 3-6 Informant number.

Col. 7 Sex of informant.

Col. 8-9 Age of informant.

Col. 10-11 Birthplace of informant.

Col. 12-13 Occupation of informant.

An input formant of A2,A4,I1,3I2 is used for the Data Identification Card.
Sample:

YE0025

(group data)

YE01001621809

(single informant)

(4) Linguistic Word Data Cards One data card is used for each linguistic word.

Col. 1-2 Language code.

Col. 3-6 Informant number.

Col. 7-9 Blanks.

Col. 10-12 Page number from test sheet.

Col. 13-15 Item number on test page. (refers to the word being phrased).

Col. 16-80 Linguistic coded word. NOTE: the maximum allowable length for the linguistic word is 65 characters.

An input format of A2,A4,T10,A3,A3,65A1 is used for Linguistic Word Data Cards.
Sample:

YE0100ßßß001002FFR$U28I02ST

(a ß denotes a blank)

(5) Key Letter Data Cards One data card is used for each key letter.

Col. 1-2 Number of characters in the Key Letter.

Col. 3-52 Key letter(s). NOTE: the maximum allowable length for the key letter(s) is 50 characters.

An input format of I2,50A1 is used for the Key Letter Data Cards.
Sample:

02R$

SAMPLE DATA DECK

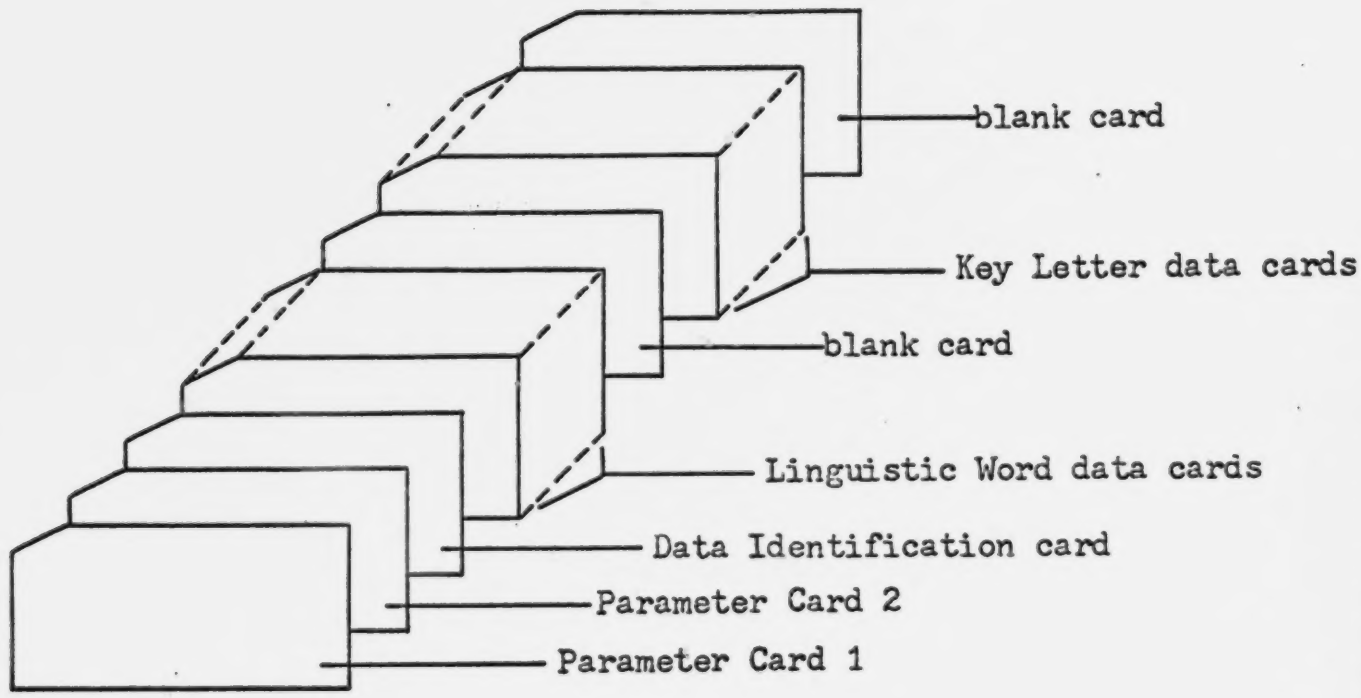
Single set of data:

Col. 1
'

0544	Control Card
013110	Parameter Card
YE01001621809	Data ID Card
YE0100ßßß001001FFR$U14....	Linguistic Word Data Cards
YE0100ßßß001002BU21I04RR	"
.	"
.	"
.	"
(ßlank card)	indicates end of data
01R	Key Letter Data Cards
02RR	"
.	"
.	"
.	"
(ßlank card)	indicates end of data

(See figure A)

FIGURE A


blank card
Key Letter data cards
blank card
Linguistic Word data cards
Data Identification card
Parameter Card 2
Parameter Card 1

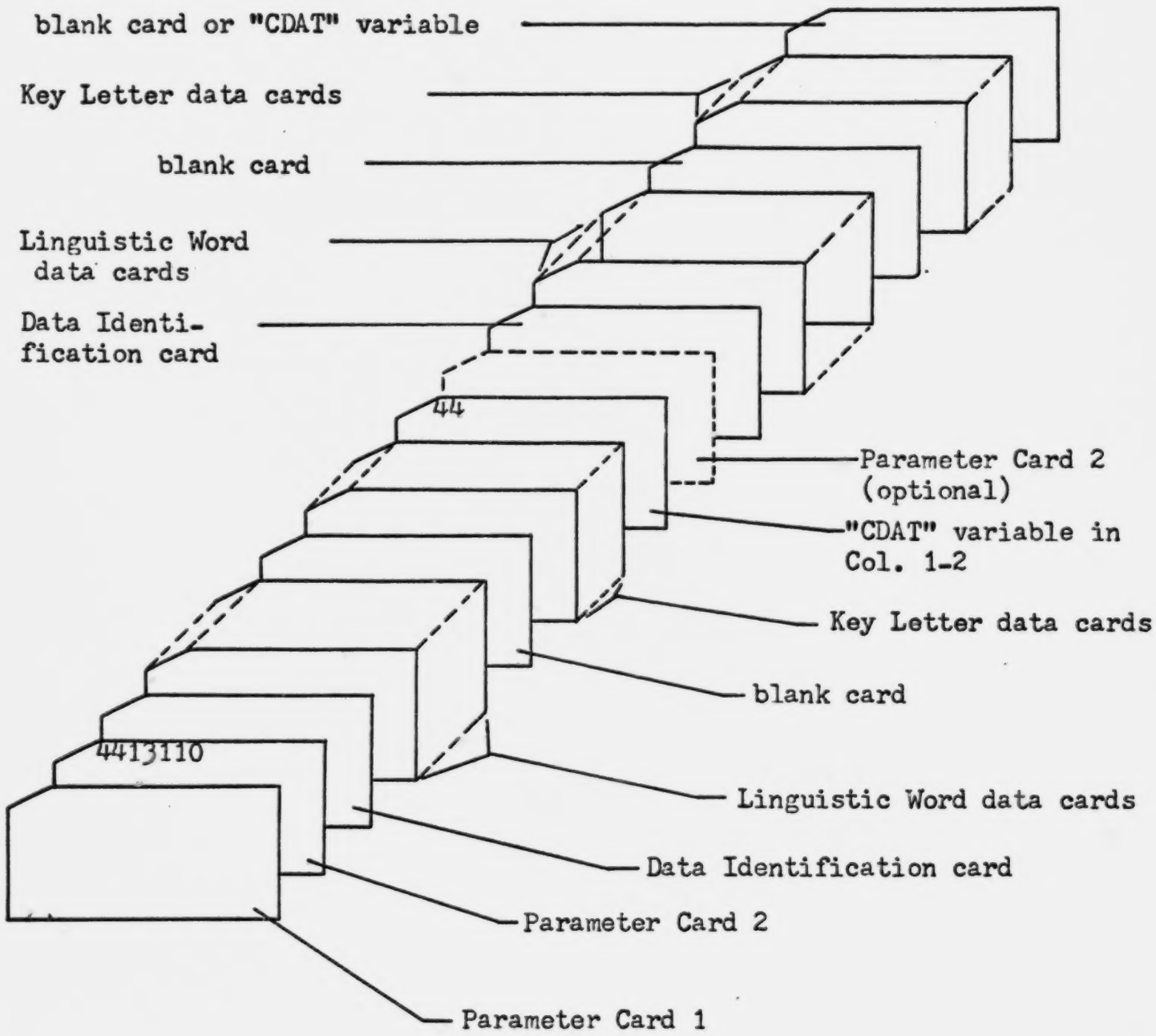
SAMPLE DATA DECK

Multiple sets of data:

```
Col. 1
'
0544                                  Control Card
013110                                Parameter Card
YE01001621809                         Data ID Card
YE0100ØØØ001001FU16RR...              Linguistic Word Data Cards
YE0100ØØØ001002TTU14I02....
.
.
.
(Ølank card)                          indicates end of data
01R                                   Key Letter Data Cards
02RR
.
.
.
44                                    CDAT variable-indicates more data
112110                                new Parameter Card
01R                                   Key Letter Data Cards
02RR
.
.
.
44                                    CDAT variable
012110                                new Parameter Card
YE01011551403                         new Data ID Card
YE0101ØØØ001001THU23...               new Linguistic Word Data
YE0101ØØØ001002GXU06I14...
.
.
.
(Ølank card)
02TH                                  Key Letter Data Card
02RX
.
.
.
(Ølank card)                          indicates end of data
```

(See figure B)

FIGURE B



Note:

1. A "blank"card always follows the set of Linguistic Word data cards.

2. A "CDAT" variable card, as illustrated above, always follows the Key Letter data cards if another set of data is to be processed.

3. A "blank" card always follows Key Letter data cards if no more data is to be processed.